DOCUMENT RESUME

ED 157 932 TM 007 311

AUTHOR Perkins, Elizabeth Phyfe; Swaminathan, Hariharan
TITLE Data Collection and Analysis in the Naturalistic Evaluation of Early Childhood Programs.
PUB DATE Mar 78
NOTE 33p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Behavior Rating Scales; Classroom Environment; *Classroom Observation Techniques; *Data Analysis; *Data Collection; Early Childhood Education; *Environmental Influences; *Evaluation Methods; Hypothesis Testing; Interaction Process Analysis; Program Evaluation; Social Behavior; *Test Reliability
IDENTIFIERS Behavioral Map (Ittelson Rivlin and Proshansky); Behavior Checklist (Perkins and Swaminathan); *Naturalistic Observation

ABSTRACT

The Naturalistic Evaluation Methodology presented here rests on a transactionist theory of child development, an ecological view of human behavior and the quasi-clinical method of inquiry. It is intended that the procedures will lead not only to the improvement of the quality of preschool education but will generate knowledge about children's behavior as a function of their interaction with the environment. Two levels of analysis have been presented that facilitate the interpretation of the data. The first, which is descriptive, involves the visual inspection of the two Behavioral Maps completed after the Behavioral Checklist has been used. (Both are included). The second, which is inferential, involves testing hypotheses concerning the effects of environmental modification on children's behavior. A time series approach based on stochastic models is advocated as the most appropriate procedure for testing these hypotheses. Procedures for data collection, determination of reliability, and data analysis are presented. Time point sampling of all children throughout the total program day is advocated, along with ratings of the environment and time sampling of the teachers' behavior. Three notions of reliability are discussed: inter-observer agreement, accuracy, and stability. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED157932

Data Collection and Analysis in the Naturalistic Evaluation of Early Childhood Programs

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

E. P. Perkins

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

by

Elizabeth Phyfe Perkins
Hariharan Swaminathan

School of Education
University of Massachusetts
Amherst, MA 01003

TM007 311

Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

The Naturalistic Evaluation Theory proposed in a companion paper (Day, 1978) assumes that children's behavior is a function of their interaction with their environment. Hunt (1975, p. 217) states:

> . . .B = f (P, E), or Behavior is a function of the Person and the Environment. . .A B-P-E analysis of a psychological experiment would specify the Behavior (or dependent variable), viewing it as jointly determined by the Person (kind of subjects) and the Environment. . . .

Application of this Behavior-Person-Environment paradigm to Early Childhood settings has resulted in the determination of seven generic categories of children's behavior as the crucial variables in the underlying equation (Day, 1978). These behaviors are expected to change as a result of changes in children's interaction with the classroom environment. The environment is conceptualized in terms of three factors: (1) the arrangement of the physical space, (2) the display and availability of materials and (3) the adult behavior.

Given this Behavior-Person-Environment paradigm the evaluation is aimed at recording and understanding behavior as it occurs in the natural setting. Measurement procedures that involve the observation of children during the life of the classroom and a rating of the environment as it is experienced by the children have been developed, i.e., both the behavior of the children and the behavior setting have to be measured.

A transactionist theory of child development forms the basis for our concern with the behavior of children at the moment. The best index of development is a measure of the activity of children as they interact with the world around them. Quantifying their performance on achievment or intelligence tests in antithetical to the naturalistic evaluation theory (Day, 1978).

Another aspect of the evaluation involves data collection and analysis by early childhood practitioners. Not only do the teaching staff participate fully in the design and operation of the procedures but also the evaluator becomes directly involved in the life of the classroom. As Fisher and Berliner (1977, pp. 14-15) have described the quasi-clinical approach:

> Quasi-clinical inquiry generates knowledge about the phenomenon while those engaging in the process [of evaluation] work cooperatively on the solution of specific classroom problems.

Given the constraints of observational techniques and a cooperative "hands-on" philosophy of evaluation, procedures for data collection and analysis have been outlined that are not only practical for use by teaching staff but also rigorous enough for reliability estimation and validation.

The primary questions which this paper addresses are: (1) Can children's behavior that needs modification be identified? and, (2) Does the manipulation of the pre-school environment according to certain theoretical principles result in developmentally appropriate changes in children's behavior? To this end we offer a set of procedures for the Evaluation of Early Childhood Programs. The Methodology will be discussed in three sections: The data collection process, determination of the reliability of the procedures, and data analysis techniques.

The Data Collection Process

The Behavior Checklist (Table 1) is the major measurement instrument. It consists of 30 discrete behaviors organized in seven generic categories. The categories have been defined as the criterion variables that characterize quality early childhood programs. Observers record the incidence

BEHAVIOR CHECKLIST DATA SHEET (3/78)

Date:_____ Time:_____

Observer_____

Program_____

	Children											Totals	Comments
Learning Area													
Group Size													
Adult Role													
TASK INVOLVEMENT: Focuses on task: materials													
TASK INVOLVEMENT: involved: no materials													
Observes intently													
Resolves problem													
Completes task, activity													
Leaves task unfinished													
Inattentive, uninvolved: sta.													
Wandering: inattentive													
Seeks participation (C/A)													
Is involved with (C/A)													
COOPERATION: Accepts adult request													
COOPERATION: Rejects adult request													
COOPERATION: Accepts child's request													
COOPERATION: Rejects child's request													
Takes turns: w or w/o mat.													
Shares: give & take													
AUTONOMY: Chooses to participate grp.													
AUTONOMY: Chooses not to participate													
AUTONOMY: Chooses activity													
AUTONOMY: Asks permission before													
COMMUNICATION: Initiates conv., req. info (A/C)													
COMMUNICATION: Engaged conversation (A/C)													
COMMUNICATION: Responds to adult, child (A/C)													
COMMUNICATION: Verbal thought-speaks to self													
COMMUNICATION: Makes irrelevant comment													
MATERIALS: Combines materials effectively													
MATERIALS: Abuses/misuses materials													
MANAGEMENT MAINTENANCE: Management, Maint. self-init.													
MANAGEMENT MAINTENANCE: Adult initiated													
MANAGEMENT MAINTENANCE: Waits, stands in line													
CONSIDERATION: Offers help/sympathy													
CONSIDERATION: Respects phys. space, materials													
CONSIDERATION: Disturbs activity, conversation													
CONSIDERATION: Teases													
CONSIDERATION: Threatens, strikes: verb/nonverb													
Other													

Table 1

of each behavior and the learning area in which the child is located. The Checklist constitutes the primary source of information for identifying children's behaviors that need modification and for inferring the effects of environmental adjustments. It is important to note that the Behavior Checklist represents behaviors which are commonly found in preschool settings and which reflect salient aspects of child development.

Observation Format and Sampling Procedures

Point-time sampling of all the children in a program can be thought of as the most appropriate data collection procedure in the present context. Point time procedures specify a set interval for observation followed by an interval for coding the behaviors noted. The recorder spends thirty seconds observing one child, followed by thirty seconds of coding. S/he then proceeds to the next child until all children have been sampled. This sequence constitutes one time point. Time points are scheduled throughout the program day.

Point-time sampling is considered appropriate for several reasons, foremost of which is the fact that it permits the sampling of all children at all times relatively easily. The method ensures that the behaviors measured are representative and hence enhances the validity of the data collected. This procedure, of brief repetitive sampling of subjects, has been found to yield more accurate and representative data on the subjects' behavior than the procedure of time sampling over long contiguous observation periods (Thomson, Holmberg, Baer, 1974). In addition, time-point sampling provides for the generation of many data points with a conservative expenditure of observer time.

However, there are two drawbacks with point-time sampling. Moderate frequency behaviors may be over-estimated when compared with high frequency behaviors of similar duration when the interval is long (Sackett, 1978). Thirty seconds is fairly long when compared with six second intervals used in other studies (Weinstein, 1977). Nevertheless, shorter intervals lose the continuity and meaningfulness that can be obtained in ten minute or longer time samples. Despite these drawbacks, time-point sampling of all children can be thought of as the most efficient, accurate and valid means of generating a representative picture of the behavior of the total group.

The Behavior Checklist, also includes the measurement of certain environmental variables. The observer indicates in what learning area or in what group activity each observation is made. An abbreviation indicating the area such as blocks, circle , painting, etc. is recorded for each child sampled, so that one can relate the observation to a particular segment of the environment. Noting, for example, "free play" would be insufficient for the subsequent analysis of the children's behavior seen as a function of their interaction with their environment.

The observer also codes two other setting variables on the Behavior Checklist: the size of the group in which the child is observed and the role of the adult nearest to the child. Group size is coded as an integer: 1, if the child is alone, 2, 3, 4 or however many children are involved . Adult behavior, is coded on a four point scale: 4, directing the activity; 3, participating; 2, watching/helping; 1, being uninvolved (Berk, 1976). Though these data are valuable they do not constitute a complete measure of the environment. Ecological specialists have been

developing ways to assess learning environments (Gump, 1974). Currently, we have been using a statement of Principles of Organization for the three environmental factors of physical space, materials presentation, and adult role/behavior. However, it is expected that ratings will be made on the various principles underlying two of the factors--space and materials. Observers will rate such aspects as the clarity of paths, the justaposition of learning areas, provision for private space, etc. The factor, adult role/behavior, partially measured by the Behavior Checklist, will be measured by an observation checklist. Observers will use a time sampling procedure of ten minute periods divided into thirty second intervals in which the frequency of salient behavior is recorded. These variables may then be employed as independent variables to obtain a measure of environment. Instrumentation for the assessment of these variables is still in progress.

The Behavioral Checklist data is conveniently summarized on a Behavioral Map (Ittelson, Rivlin, and Proshansky, 1970). The Behavioral Map, Table 2, provides a vivid picture of what kinds of behaviors occur in which learning areas. Proportional scores that indicate the frequency of each behavior by learning area facilitate comparisons of behavior across learning areas as well as within a learning area.

The Evaluation Methodology calls for two major observation periods interrupted by an environmental adjustment phase. A summary of the procedures is included at this point to give the reader an overview of the total process within which measurement activities occur.

3/78

BEHAVIORAL MAP #_____

Dates:_____

Program:_____

Table 2

		Learning Area														Totals
	Group Size: 1 / 2-4 / 4-8 / over 9															
	Adult Role: 1 / 2 / 3 / 4															
	Total Number of Children															
TASK INVOLVEMENT	Focuses on task: materials involved; no materials															
TASK INVOLVEMENT	Observes intently															
TASK INVOLVEMENT	Resolves problem															
TASK INVOLVEMENT	Completes task, activity															
TASK INVOLVEMENT	Leaves task unfinished															
TASK INVOLVEMENT	Inattentive, uninvolved: sta.															
TASK INVOLVEMENT	Wandering: inattentive															
COOPERATION	Seeks participation (C/A)															
COOPERATION	Is involved with (C/A)															
COOPERATION	Accepts adult request															
COOPERATION	Rejects adult request															
COOPERATION	Accepts child's request															
COOPERATION	Rejects child's request															
COOPERATION	Takes turns: w or w/o mat.															
COOPERATION	Shares: give & take															
AUTONOMY	Chooses to participate grp.															
AUTONOMY	Chooses not to participate															
AUTONOMY	Chooses activity															
AUTONOMY	Asks permission before															
COMMUNICATION	Initiates conv., req. info (A/C)															
COMMUNICATION	Engaged conversation (A/C)															
COMMUNICATION	Responds to adult, child (A/C)															
COMMUNICATION	Verbal thought/speaks to self															
COMMUNICATION	Makes irrelevant comment															
MANAGEMENT MATERIALS	Combines materials effectively															
MANAGEMENT MATERIALS	Abuses/misuses materials															
MAINTENANCE	Management, Maint. self-init. / Adult initiated															
MAINTENANCE	Waits, stands in line															
CONSIDERATION	Offers help/sympathy															
CONSIDERATION	Respects phys. space, materials															
CONSIDERATION	Disturbs activity, conversation															
CONSIDERATION	Teases															
CONSIDERATION	Threatens, strikes: verb/nonverb.															
	Other															

Summary of the Data Collection Process

Phase I. All children are observed on a time-point sampling basis throughout the program day for 10-20 days. Behavior totals for each learning area are arranged on a Behavioral Map. Data is gathered on the physical setting, the materials arrangement and the adult role. Behavior totals for each learning area are arranged on a Behavioral Map.

Phase II. The teacher and clinician examine the Behavioral Map to identify learning areas where the intended behavior of children does not match the obtained behavior. Measures of the environment are examined for indications of environmental variables that are amenable to modification. Changes in the preschool setting are implemented and measured. Desired changes in children's behavior in each learning area are predicted. The children are observed during an adjustment period of several days.

Phase III. Observations on all children and measurement of the environment are carried out according to a schedule identical to Phase I. Summary data of all behaviors are arranged by learning area on a second Behavioral Map.

The data collection procedures outlined are designed for use by early childhood practitioners. These procedures have been chosen because of their appropriateness for use in natural settings and on account of the multivariate picture of children's behavior by learning area that emerges. However, before these data can be meaningfully interpreted, the reliability of the procedures in quasi-clinical settings has to be established.

Reliability of the Observation Procedures

The importance and complexity of reliability estimation for observational measures has been discussed by Medley and Mitzel (1963), Johnson and Bolstad (1973), Rowley (1974), Herbert and Attridge (1975) and Hollenbeck (1978) among others. Reliability is defined as a characteristic of a procedure for quantifying characteristics of individuals, and hence, is situation dependent. As noted by Rowley (1974, p. 53),

> A single instrument may produce scores which are reliable, and other scores which are unreliable.. Even one measure may be reliable or unreliable depending on the manner in which the instrument is used, the subjects observed, the skill of the observer and the number and length of observation periods.

This observation is particularly relevant for data gathered in quasi-clinical settings. Hence, before such data can be analyzed and interpreted, it is crucial to establish the "reliability" of the data collection process. In the following discussion, reliability is conceptualized in terms of three components: interobserver agreement, accuracy and stability.

Interobserver Agreement

Interobserver agreement is defined as the percentage of agreement between two observers, the evaluator and the teacher, coding the same subjects at the same time. It is the most frequently reported reliability statistic in the observational literature. However, Medley and Mitzel (1963) and McGaw, Wardrop and Bunda (1972) have argued cogently against the adequacy of observer agreements as the sole index of reliability. Nevertheless, achieving high interobserver agreement is a prerequisite to establishing the accuracy and stability of the procedures.

Observer agreement is often computed in the following manner: the number of agreements divided by the number of agreement plus disagreements.

However, since Hollenbeck (1978) notes at least six other indices, it is important to choose the index that is most appropriate to the research questions being asked. In addition, Johnson and Bolstad (1973) advocate the procedure of comparing the obtained interobserver agreement with that expected by chance alone. Such a comparison is important for the interpretation of the size of the agreement coefficient.

At the moment two indices appear appropriate for the assessment of interobserver agreement. A procedure presented by Bernstein (1974), although intended to yield an estimate of accuracy gives, the probability of two or more observers agreeing. Adapting Bernstein's method, the interobserver reliability can be defined as

$$P_{xy} = \sqrt{A_{xy}}$$

where

$$A_{xy} = \frac{\text{number of agreements between observers x and y}}{\text{number of agreements plus disagreements}}$$

However, the index, P_{xy}, does not take into account agreement that can be expected by chance alone.

The kappa coefficient, κ, (Swaminathan et al., 1974) takes chance agreement into account and hence, is more appropriate. In order to compute κ, the data is arranged as shown in Table 3.

Table 3

Observer I / Observer II	Behavior Observed	Behavior not Observed	Marginal Proportions
Behavior Observed	P_{11}	P_{10}	$P_{1.}$
Behavior not Observed	P_{01}	P_{00}	$P_{0.}$
Marginal Proportions	$P_{.1}$	$P_{.0}$	

The quantity κ is defined as

$$\kappa = (P - P_c)/(1 - P_c),$$

where, P, the observed proportion of agreement, is given by

$$P = \Sigma \; P_{ii} = P_{11} + P_{00}$$

and, P_c, the proportion of agreement expected by chance alone is given by,

$$P_c = \Sigma \; P_{i\cdot} \, P_{\cdot i}$$

$$= P_{0\cdot} \, P_{\cdot 0} + P_{1\cdot} \, P_{\cdot 1}$$

When more than two observers are available, computation of interrater reliability using κ becomes tedious. In this case an Analysis of Variance procedure may be employed. In ANOVA terms interobserver reliability is defined as:

$$\frac{\text{variance due to observers}}{\text{variance due to observers + variance due to error}}$$

In terms of classical test theory, this index can be taken to reflect the "internal consistency" of the raters.

A high degree of interrater reliability is clearly a prerequisite for data collection. In addition, raters should be calibrated against each other throughout the study in order to maintain the desired level of agreement. It is important to note that the index of agreement has little bearing on the significant issues of accuracy and stability.

Accuracy

Since the Behavior Checklist, presented earlier, constitutes the only measure of the dependent variable, it is crucial that the accuracy of the data collection procedures be established. The notion of accuracy deals with the degree of agreement between an observer's record of behavior and a criterion record of the same behavior. Hence, in a sense,

the issue of accuracy is closely related to the determination of the validity of the procedures in natural settings.

The basic problem in determining the accuracy of observational methods lies in establishing a suitable criterion protocol with which to compare observers records. Two procedures are available for this purpose. One method is to have an "expert" observer record the behavior of children in a classroom at the same time as another observer. The protocol of the expert is then used as the criterion against which the other observer's record is compared. The other method is to have experts code videotaped segments of children's behavior until they agree on a criterion protocol. The observers then code the same segment. In either case, the kappa coefficient (Swaminathan et al., 1974) or the index given by Bernstein (1974) can be taken as the measure of accuracy.

Stability

In the Naturalistic Evaluation Procedure, stability is defined as the consistency of children's behavior over time within learning areas. With reference to this notion, Medley and Metzel (1963) point out that the consistency of behavior measured by different observers at different times, constitutes the true measure of reliability. This notion of stability is crucial to any attempts to validate the evaluation procedures and their content. For the Behavior Checklist to yield meaningful data about the dependent variable, the behaviors as measured must demonstrate consistency over time. Otherwise, attempts to modify such behavior cannot be empirically tested, since a large amount of variation may obscure the effects of any environmental adjustment. In order to facilitate the detection of such effects, stability will be obtained over a period of weeks during Phase I data collection.

In order to assess stability, the data may be conveniently arranged as shown below:

Table 4

	D_1	D_2	. . .	D_j	.	D_p
O_1	X_{11}	X_{12}		.		X_{1p}
.				.		
.				.		
.				.		
O_i	X_{i1}	X_{i2}	. .	X_{ij}	.	X_{ip}
O_k	X_{k1}	X_{k2}				X_{kp}

where X_{ij} is the score for the ith Observer on the jth Day (i=1, ..., k; j-1, ..., p). In order to compute the score X_{ij}, we assume that there are t time points within a day, and that f_ℓ is the total frequency of behavior observed at time point ℓ (ℓ=1, ..., t), and n_ℓ is the number of children observed in the learning area at time point ℓ. With this, X_{ij} is given as

$$X_{ij} = (\sum_{\ell=1}^{t} f_\ell) / \sum_{\ell=1}^{t} n_\ell$$

The reliability coefficient r_s is defined as:

$$r_s = \frac{MS_D - MS_W}{MS_D + (p-1)\ MSW}$$

A high value for r_s indicates stability.

Such stability coefficients can be computed for each generic category within each learning area. In typical classroom situations, where the number of observers range from two to three, stability would rarely be calculated. However, stability must be established in quasi-clinical settings, involving both evaluators and teachers, before the procedure may be assumed to yield reliable data.

Factors Affecting Reliability

Observer bias. The obstrusiveness of observers in naturalistic setting may influence the behavior of the subjects. However, Medley and Mitzel (1963) state that children soon adapt to the presence of an observer in the classroom. Johnson and Bolstad (1973) support this point of view:

> The literature also suggests that age is correlated with reactivity, several authors (Barker and Wright, 1955; Polansky et al., 1949) have suggested that younger children are less self-conscious and thereby less subject to reactive effects than older children.

The obtrusiveness of the observers given the population studied (e.g., 2 years 9 months to 5 years), is expected to have minimal impact on the validity of the data collected.

Observer bias and expectancy effects are unavoidable threats to the reliability of quasi-clinical evaluation procedures. These sources of systematic error may easily be present becuase the evaluator and the teacher are aware of the nature of the study and the desired outcome of the environmental modifications. Observer bias is defined as errors in the coding protocol reflecting either behaviors that have not occurred or omission of behaviors that are present but are not coded.

An additional consideration in assessing the potential contributions of such observer bias is the degree of specificity of the behavior code. The more molar or global are the categories to be coded, the more room there is for differential observer interpretation and bias. Johnson and Bolstad (1973, p. 37) state that ". . .a multivariate coding system with well-defined behavioral codes might be expected to restrict interpretive bias. This is an empirical question worthy of examination." Since the Behavior Checklist involves seven functional areas represented by 30 behaviorally defined categories, it is assumed that adequate training will reduce this type of observer bias.

The fact that the evaluator and the teaching staff are intimately involved with both the data collection and the treatment design, not only allows for the possibility of bias but also raises the issue of observer expectancy effects. The fact that a teacher may communicate his/her expectations to the children (Rosenthal and Jacobson, 1966) and thus influence their behavior is a legitimate function of the adult role in the classroom. However, when the teacher observes the children s/he must attempt to be neutral about their behavior. Nevertheless, it is most likely that expectancy effects are in operation during teacher data collection. To a lesser degree they may also operate during coding by the evaluators.

The Unit of Measurement. As Johnson and Bolstad (1973) have pointed out, the unit of measurement used in interobserver and accuracy estimation should be the same unit score on which the analysis of the dependent variable is performed. Without such a restriction, reliability coefficients may not truly estimate the accuracy and reliability of the scores used to estimate effects. Calculations of interobserver agreement and accuracy will utilize the frequency of a behavior summed across all children recorded in a learning area at one time point which is the smallest unit score employed in the data analysis.

Summary

Three components of reliability have been discussed in terms of their meaning for observational purposes, their importance in naturalistic evaluation and their estimations in quasi-clinical settings. Both observer bias and expectancy effects have been identified as potential threats to internal validity. The third factor, reactivity of children to observers, is expected to have little effect on reliability

due to the age of the children. The appropriate unit of measurement for reliability estimation is the unit of measurement that is employed in the data analysis phase.

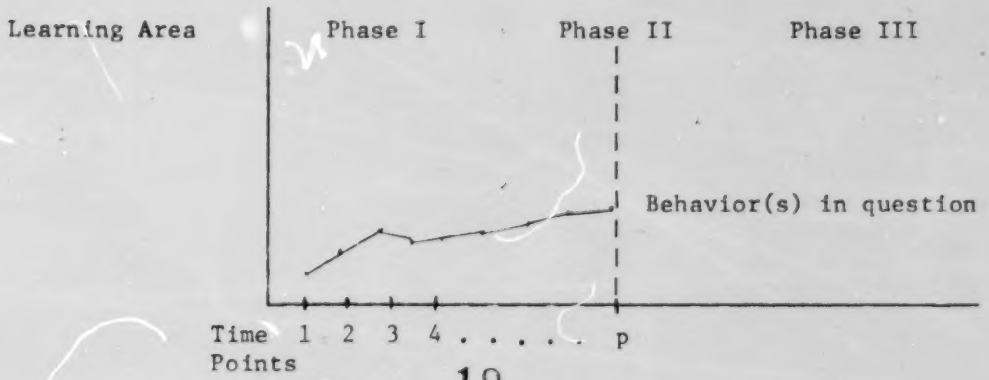
It is recommended that interrater reliability be computed throughout the data collection period. Observer accuracy should reach a criterion level before data collection begins. Although stability seldom needs to be checked as a part of the ongoing data collection process, it must be established prior to the general application of the procedures. As pointed out earlier, the three notions of reliability, interobserver agreement, accuracy and stability, are necessary for establishing the validity of the overall procedures and for inferring effects due to environmental adjustments.

DATA ANALYSIS

The analysis of the data collected in early childhood settings can be conceptualized on two levels. The first, which is descriptive, involves data summary procedures and guidelines for analysis-by-inspection at the local level. A description of these procedures involving the program staff and the evaluator are presented first. The second level of analysis, which is inferential, consists of the utilization of statistical procedures for inferring the effects of environmental adjustments. Both analyses are complementary aspects of the quasi-clinical approach to early childhood evaluation: the first allows for immediate feedback and intuitive analysis of the data by program staff at the classroom site; the second permits testing of hypotheses concerning the major research questions and serves to validate the behavior-person-environment interaction theory.

Level I: Analysis by Inspection

The first evaluation issue with which the evaluator and the teacher are concerned at the local level is the determination of the need for environmental modification. At the end of the Phase I Data Collection period, the evaluator and the teacher summarize the data on the Behavioral Map (Table 2). The evaluator may also construct graphs of particular behaviors by learning areas for further clarification, as shown below:



The initial step in the analysis involves the comparison of the obtained behavior profile with the intended behaviors for each learning area. The program staff obtains an indication of the degree to which they have met the goals specified previously for each learning area. They identify behaviors whose frequency they wish to increase or decrease and, they establish specific expectations for behavioral change in terms of each learning area. This informal process rests on their previous comparison of intended and obtained behaviors as well as on any intuitive reassessment of goals brought about by observing and recording the children's behavior. For example, the staff may decide they wish to raise task involvement in the reading area. In an examination of the environmental data, they may notice that the block area, with all its noisy boisterous play, is located next to the reading corner. They also may note that the teacher role is recorded primarily as "direct leadership." Bringing to bear all their familiarity and experience with the particular early childhood setting, the evaluation team looks for relationships between such environmental factors and the obtained children's behavior. At this point the staff may design the following environmental modification. The reading corner would be moved to a quiet section of the room, equipped with pillows and a warm rug. A low book shelf may be arranged to partially screen off the area from the rest of the room. It would be planned that adults occasionally enter the area to read to a child or engage in discussion about something in a book.

The next step involves the prediction of changes in children's behavior that are expected as a result of the adjustments in the environment. The purpose of such predictions is not to test local program staff but to encourage the exploration of the behavior-person-environment paradigm in early childhood settings.

The program staff implements all environmental modifications in Phase II of the Evaluation Procedure. Those that involve changes in adult behavior are written out as specifically as possible in terms of daily classroom practice. After a brief adjustment phase in which the children are observed in order to detect extreme behavioral shiftsdue to the novelty of the change, the previous schedule of data collection is repeated.

The evaluator and program staff must now address the second basic evaluation issue, that of analyzing the effects of the modifications in the environment. A second Behavioral Map is prepared so that learning area profiles can be compared with the previous profiles. The evaluator completes any graphs of behaviors begun in Phase I. The evaluation team carries out a comparison of predicted vs. obtained changes in children's behavior by learning area. An inspection of the graphs of the behaviors over time by learning area is suggested if the evaluation lasts fifteen weeks. In that length of time, a growth trend in a behavior may be mistaken for a true change in behavior. A graph of the behaviors may suggest a progression to the second level of analysis.

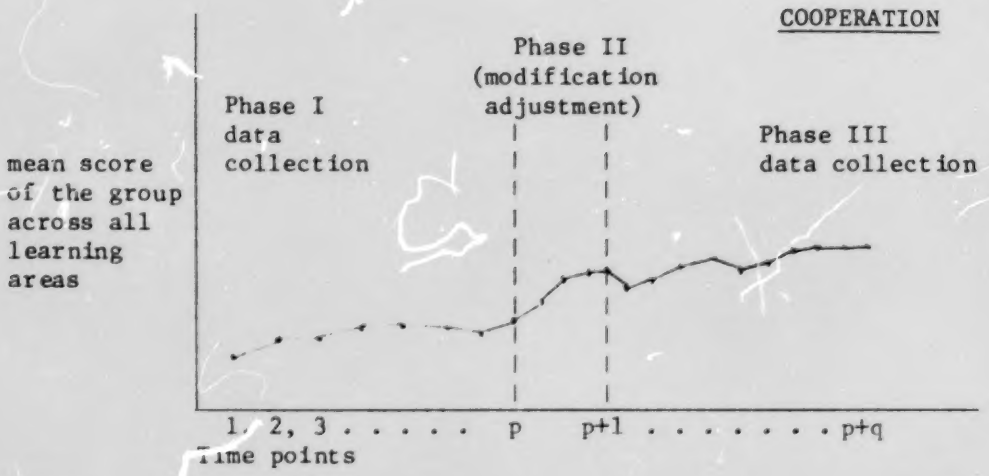
Level II: Statistical Analysis

The second level of analysis, is designed to answer some of the more ambitious questions generated by the application of the evaluation procedures. This involves determining changes in children's behavior over and above the effects of a rapid developmental trend. In addition, we may seek to verify environmental effects assumed to be significant at the local level. Whereas the first level seeks to provide informal evaluation procedures that are immediately available to early childhood

program staff, the second level of analysis provides for the ongoing validation and empirical testing of the environmental modifications.

The basic research question at this level concerns the inference of effects of environmental modifications on the behavior of children in a particular pre-school seeting. Since, we assume that children's behavior is a function of children's interaction with their environment, we can expect the planned shifts in the three environmental variables to result in changes in the children's behavior. In order to obtain indications of such change, the two behavioral maps, one summarizing children's behavior before the modification and the other representing behavior after the modification, may be inspected. However, statistical analysis of the data is necessary in order to infer the effects of the modifications. Changes in the children's behavior on each of the seven generic categories (e.g., cooperation, autonomy, communication, center maintenance, consideration and task involvement), can be stated in terms of five major research questions.

The most general question involves an analysis of the children's scores on each of the behavior clusters. Have there been overall shifts in their behavior? The most appropriate unit of analysis for studying this question is the average frequency of the specific behavior in the particular category collapsed across all children and all learning areas for one time point. For example, suppose that at the end of phase I, a teacher decides that s/he wishes to raise the level of cooperative behavior of the preschool class. Environmental modifications are planned and carried out in phase II. Data is collected during phase III. The following hypothetical graph may be drawn from the data:



The question to be answered is: Has the group's general performance on particular variables selected by the evaluation team changed significantly from Phase I to Phase III?

Since many of the environmental adjustments are made in terms of "problem" areas of the classroom, it is crucial to ask whether children's behavior has altered within those areas. Suppose the block area has revealed a negative score on consideration and a low score on task involvement during Phase I. The teacher and clinician examine the physical space, the teacher role and the arrangement and nature of the materials. Alterations in the block area are planned in order to raise task involvement and consideration behaviors. The unit of analysis in this case is the mean score for the category (e.g., task involvement or consideration), of all children coded in the block area at each time point. The data may be graphed in a manner similar to the example given above. The question to be answered is: Has the children's behavior in the specified area altered significantly from Phase I to Phase III for a generic category of behavior?

The third question is: Are there changes in the children's use of each learning area. For example, a teacher may wish to raise the participation rate of children in the table activity area after discovering a low density rate for the area in the Phase I data. The unit of analysis for this question is the number of children recorded in the table activity area at each time point.

The fourth question concerns the analysis of changes in the percentage of children scored as playing individually, in pairs, in groups of 3, 4, 5, etc. The fifth examines shifts in the record adult role. These analyses may be of interest as checks on the implementation of the modification planned by the staff. Alterations in adult role/behavior may have been planned in order to change group size so that a behavior category may be increased. For instance, in Phase I, if children are found mainly in large groups or by themselves, instances of cooperative behavior may be expected to be quite low. Adults may plan to take on the role of participator more frequently in order to stimulate small group interaction and thus, hopefully, cooperative behaviors. Group size in this case may be thought of as a mediating variable, and adult role/behavior as an independent variable. The unit of analysis for studying the changes in group size is the proportion of children playing individually in 2's, 3's, etc., at each time point. For the adult role, a category label assigned to each role, is recorded for each child. A proportional score for each role at each time point is then computed as the unit of analysis.

All of the above questions can be thought of in terms of a time series research design. Observations of all the children in an early childhood program are conducted on a point time sampling basis, such that all times of the program day are sampled equally. The research design with (p+q) observations may be diagrammed as follows:

$O_1 \quad O_2 \quad O_3 \cdots O_p \qquad X \qquad O_{p+1} \quad O_{p+2} \cdots O_{p+q}$

Phase I Phase II Phase III

where O_i designates an observational unit at time point i $(i = 1, \ldots p+q)$, and X designates the environmental modification and short adjustment phase.

Naturalistic research, by definition, seeks to generate knowledge about behavior in the natural settings. Thus, traditional experimental practices of controling the equality of settings and the self-selection of subjects are antithetical to the purposes of the evaluation methodology. The advantages of time series design for situations in which the experimenter has little control over the milieu or the assignment of subjects has been summarized by Campbell and Stanley (1963). Thus, structuring the data collection as a time series facilitates the inference of the effects of environmental manipulation in early childhood classrooms. However, in keeping with the quasi-clinical approach to evaluation of Fisher and Berliner (1977), we are primarily concerned with threats to the internal validity of this design and not external validity.

The time series design controls for all rival hypotheses except that of history and hence, is a sound design for testing hypotheses concerning the changes in children's behavior following adjustments in the physical space, materials display or adult role/behavior. We suggest several precautions in carrying out the evaluation procedures in order to eliminate history as a plausible rival hypothesis. The usual curriculum should not be altered and pertinent outside events should be recorded throughout the evaluation period. Examination and logical analysis of the influence of such events will then preclude history as a threat to the internal validity of the design.

Campbell and Stanley (1963) review possible tests of significance for inferring an intervention effect in the time-series design. They point out that the comparison of O_p with O_{p+1} does not provide an adequate test of the environmental effect since a difference between O_p and O_{p+1} may be the part of a developmental trend.

Similarly, a comparison of the mean of all the pre-adjustment observations with the mean of all the post-adjustment observations does not provide a valid test of the adjustment effect since, in the presence of a developmental trend, the two means could be different without an adjustment effect. The usual analysis of variance procedure for comparing means cannot be employed to test the effect of the adjustment since these procedures require the observations across time to be independent, an untenable assumption in the present situation.

Multivariate procedures for inferring an adjustment effect that take into account the dependence in the data have been outlined recently by Algina and Swaminathan (1977) and Swaminathan and Algina (1977). However, these procedures require that the number of individuals measured at each time point exceed the number of time points and hence may not be appropriate in naturalistic studies.

If x_t is a sequence of random variables measured at times $t = 1, 2, 3, \ldots, p, p+1, \ldots, p+q$, and e_f is the error associated with the variable x_t, then an ARIMA model of order (p, d, q) can be represented as

$$f(B)\ \Delta^p\ x_t = g(B)\ e_t\ .$$

Where B and Δ are linear operators such that $Bx_t = x_{t-1}$,

$$\Delta x_t = x_t - x_{t-1} = (1-B)\ x_t\ ,$$

$$f(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$$

and

$$g(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q \quad .$$

For instance, if p = 1, d = 0, q = 0, then the ARIMA model becomes

$$x_t - \phi x_{t-1} = e_t.$$

This model, known as a first order auto-regressive model, implies that the observation at time t depends only on the immediate past. When this relationship is taken into account in this model, the errors e_t can be considered independent of each other, and hence, conventional statistical methods can be employed to analyze the data.

In particular, it can be shown that

$$x_t = \sum_{i=E}^{t} w_i e_i, \quad t = 1, 2, \ldots, p, p+1, \ldots, p+q,$$

so that each x_t is a linear combination of independently and identically distributed errors. The weights w_i are functions of $\theta_1, \theta_2, \ldots, \theta_q, \phi_1, \phi_2, \ldots, \phi_p$ and, hence, are determined once the parameters of the ARIMA model are determined. In order to evaluate the effectiveness of an intervention, Glass et al. (1975), assume that the initial level of the of the sequence x_t is L and after the intervention, the level changes to L + δ. Thus for the pre-adjustment observations,

$$x_t = L + \sum_{i=1}^{t} w_i e_i \quad , \quad t=1, \ldots, p,$$

and for the post-adjustment observations,

$$x_t = L + \delta + \sum_{i=1}^{t} w_i e_i \quad , \quad t = p+1, \ldots, p+q.$$

The hypothesis of no adjustment effect can be tested by testing the hypothesis H_o: $\delta = 0$, using the procedure given by Glass et al., (1975). This procedure has been successfully applied to evaluate the effectiveness of environmental adjustment in an early childhood education program (Weinstein, 1977).

The procedure described above is ideally suited for clinical and quasi-clinical studies since the procedure can be applied to study the behavior of a single organism or unit. However, in order to apply the procedure developed by Glass et al (1975), it is first necessary to estimate the parameters of the ARIMA model, and treat them as *known* quantities when determining the effectiveness of an intervention. Since the parameters of the ARIMA model are estimated, the estimates, are subject to sampling fluctuations. Treating estimates as parameters can be expected to bias the procedure for inferring treatment effect in an unknown way. Coulson and Swaminathan (1978) have eliminated this problem by outlining a procedure for the simultaneous estimation of the parameters of the stochastic model and the effect of the adjustment.

In essence, Coulson and Swaminathan (1978) showed that the observation at time t can be expressed as

$$x_t = \alpha_o + \alpha_1 t + e_t \qquad t = 1, \ldots, p,$$

and

$$x_t = \beta_o + \beta_1 t + e_t \qquad t = p+1, \ldots, p+q.$$

If the errors e_t (t=1, ..., p+q) are assumed to follow a first order auto-regressive process, i.e.,

$$e_t = \rho e_{t-1} + \varepsilon_t$$

where ε_t are independently and identically distributed then, then conditional model for x_t given x_{t-1} and x_{t+1} can be expressed as

$$x_t = \alpha_o^* + \alpha_1^* t + \gamma(x_{t-1} + x_{t+1}) + e_t^* \qquad t = 1, \ldots, p ;$$

and

$$x_t = \beta_o^* + \beta_1^* t + \gamma(x_{t-1} + x_{t+1}) + e_1^* \qquad t = p+1, \ldots, p+q.$$

The quantities e_t^* are independent and hence the parameters α_o^*, α_1^*, β_o^* and β_1^* may be estimated by employing the usual least squares procedures. If $\alpha_o^* = \beta_o^*$ and $\alpha_1^* = \beta_1^*$, then the hypothesis of an adjustment effect is not tenable. These hypotheses can be tested in the usual manner (Searle, 1971) and the analysis carried out with the aid of existing computer programs that carry out regression analysis. This procedure can be extended to handly a pth order auto-regressive scheme and when there is a polynomial trend. For details we refer the reader to the above authors.

Data analysis involves two complementary components. For the first component, the evaluator and program staff carry out an inspection of the Phase I and Phase III Behavioral Maps. Such inspection forms the basis for judgments about program quality and the environmental modifications. The second component involves statistical analysis of the data. The procedures discussed above are crucial for the inference of effects of such modifications upon children's behavior.

CONCLUSIONS

The procedures outlined above for the naturalistic Evaluation of Early Childhood Education Programs are based on quasi-clinical inquiry. As Fisher and Berliner (1977, p. 21) assert

> Quasi-clinical inquiry can be rigorous, experimental, and used in the service of theory construction. . . .
>
> In our view, positively valued changes in educational practice as well as generation of educational theory are the major criteria for judging a method of inquiry. Conventional research has been of very limited relevance for these purposes. We believe that quasi-clinical inquiry which addresses classroom problems, and strives to bring about change in the sites where the inquiry is conducted, has great potential for both instructional theory and classroom practice.*

Procedures for data collection, determination of reliability and data analysis have been presented in this paper. Time point sampling of all children throughout the total program day has been advocated, along with ratings of the environment and time sampling of the teachers behavior. Before the resulting data can be analyzed, reliability must be assured. Three notions of reliability were discussed: inter-observer agreement, accuracy and stability. Minimum standards must be established and met for each index of reliability. Only then can valid inference of effects can be made.

Two levels of analysis have been presented that facilitate the interpretation of the data. The first, which is descriptive, involves the visual inspection of the two Behavioral Maps by the program staff and the evaluator. The second, which is inferential, involves testing hypotheses concerning the effects of environmental modification on children's behavior. A time series approach based on stochastic models is

*Emphasis is the authors.

advocated as the most appropriate procedure for testing these hypotheses. The results of this analysis may be used to solidify the clinical findings as well as to generate further research questions concerning growth trends and specific environment-behavior interactions of children in preschool settings.

The Naturalistic Evaluation Methodology presented here rests on a transactionist theory of child development, an ecological view of human behavior and the quasi-clinical method of inquiry. It is our intention that the procedures will lead not only to the improvement of the quality of preschool education at the moment but will generate knowledge about children's behavior as a function of their interaction with the environment.

References

Algina, J. & Swaminathan, H. A procedure for the analysis of time-series designs. Journal of Experimental Education, 1977, 45, 56-61.

Berk, Laura E. How well do classroom practices reflect teacher goals? In Young Children, November 1976.

Bernstein, A. L. An estimate of the accuracy (objectivity) of nominal category coding. In The Classroom Interaction Newsletter, Vol. 4, No. 2, May 1969, 49-52.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In Gage, N. L. (Ed.), Handbook of Research on Teaching. Skokee, IL: Rand McNally, 1963.

Coulson, D., & Swaminathan, H. Analysis of single observation quasi-experimental time series designs. A paper presented at the meeting of the American Educational Research Association, Toronto, 1978.

Day, D. E. "The Development of a Naturalistic Early Childhood Education Evaluation Procedure. Paper presented at the Annual conference of the American Educational Research Association in Toronto, Canada, March 1978.

Fisher, C. W., & Berliner, D. C. Quasi-clinical inquiry in research on classroom teaching and learning. Paper presented at the meeting of the American Educational Research Association, New York, 1977.

Glass, G. V., Willson, V. L., & Gottman, J. M. Design and analysis of time-series experiments. Boulder, Col.: Laboratory of Educational Research, University of Colorado, 1975.

Gump, Paul V. Intra-setting Analysis: The third grade classroom as a special but instructive case in Willems, E., & Rausch, H. (Eds.), Naturalistic Viewpoints in Psychological Research. New York: Holt, 1969.

Herbert, J., & Attridge, C. A guide for developers and users of observation systems and manuals. American Educational Research Journal, 1975, 12, 1-20.

Hollenbeck, A. R. Problems of reliability in observational research. In Sackett, G. P. (Ed.), Observing Behavior Volume II. Baltimore: University Park Press, 1978.

Hunt, D. E. Person-environment interaction: A challenge found wanting before it was tried. Review of Educational Research, 1975, 45, 209-230.

Ittelson, W. H., Rivlin, L. G., & Proshansky, H. M. The use of behavioral maps in environmental psychology. In Proshansky, H. M., Ittelson, W. H., & Rivlin, L. G. (Eds.), Environmental Psychology: Man and His Physical Setting. New York: Holt, Rhinehart & Winston, 1970.

Johnson, S. M., & Bolstad, O. D. Methodological issues in naturalistic observations: Some problems and solutions for field research. In Hamerlynck, L. A., Handy, L. C., & Mash, E. J. (Eds.), Behavior Change: Methodology, Concepts and Practice. Champaign, IL: Research Press Company, 1973.

Medley, D. M., & Mitzel, H. E. Measuring classroom behavior by systematic observation. In Gage, N. L. (Ed.), Handbook of research on teaching. New York: Rand McNally & Co., 1963.

McGraw, B., Wardrop, J. L., & Bunda, M. A. Classroom observation schemes--Where are the errors? American Educational Research Journal, 1972, 9, 13-27.

Rowley, G. L. The reliability of observational measures. American Educational Research Journal, 1976, 13, 51-59.

Sackett, G. P. (Ed.) Measurement in observational research in observing behavior, Baltimore: University Park Press, 1978.

Searle, S. R. Linear Models. New York: John Wiley & Sons, Inc., 1971.

Soar, R. S. Final report: Follow through classroom process measurement and pupil growth (1970-71). Gainesville, FL: Institute for Development of Human Resources, University of Florida, 1973.

Swaminathan, H., & Algina, J. Analysis of quasi-experimental time series designs. Multivariate behavioral research. January 1977, 12, 111-131.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

Thomson, C., Holmberg, M., & Baer, D. A brief report on a comparison of time sampling procedures. Journal of Applied Behavior Analysis, 1974, 7, 623-626.

Weinstein, C. S. Modifying student behavior in an open classroom through changes in physical design. American Educational Research Journal, 14, 249-262.